

RESEARCH ARTICLE

# Hierarchical Novelty-Familiarity Representation in the Visual System by Modular Predictive Coding

Boris Vladimirskiy<sup>‡</sup>, Robert Urbanczik, Walter Senn\*

Department of Physiology, University of Bern, Bülhplatz 5, 3012 Bern, Switzerland

<sup>‡</sup> Current address: Munich Re UK & Ireland Life, Plantation Place, 30 Fenchurch Street, London EC3M 3AJ, United Kingdom

\* [Senn@cns.physio.unibe.ch](mailto:Senn@cns.physio.unibe.ch)



OPEN ACCESS

**Citation:** Vladimirskiy B, Urbanczik R, Senn W (2015) Hierarchical Novelty-Familiarity Representation in the Visual System by Modular Predictive Coding. PLoS ONE 10(12): e0144636. doi:10.1371/journal.pone.0144636

**Editor:** William W Lytton, SUNY Downstate MC, UNITED STATES

**Received:** January 27, 2014

**Accepted:** November 22, 2015

**Published:** December 15, 2015

**Copyright:** © 2015 Vladimirskiy et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was supported by the Swiss National Science Foundation (SNSF, personal grants No. 31003A-133094 and 310030L-156863 of WS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

## Abstract

Predictive coding has been previously introduced as a hierarchical coding framework for the visual system. At each level, activity predicted by the higher level is dynamically subtracted from the input, while the difference in activity continuously propagates further. Here we introduce modular predictive coding as a feedforward hierarchy of prediction modules without back-projections from higher to lower levels. Within each level, recurrent dynamics optimally segregates the input into novelty and familiarity components. Although the anatomical feedforward connectivity passes through the novelty-representing neurons, it is nevertheless the familiarity information which is propagated to higher levels. This modularity results in a twofold advantage compared to the original predictive coding scheme: the familiarity-novelty representation forms quickly, and at each level the full representational power is exploited for an optimized readout. As we show, natural images are successfully compressed and can be reconstructed by the familiarity neurons at each level. Missing information on different spatial scales is identified by novelty neurons and complements the familiarity representation. Furthermore, by virtue of the recurrent connectivity within each level, non-classical receptive field properties still emerge. Hence, modular predictive coding is a biologically realistic metaphor for the visual system that dynamically extracts novelty at various scales while propagating the familiarity information.

## Introduction

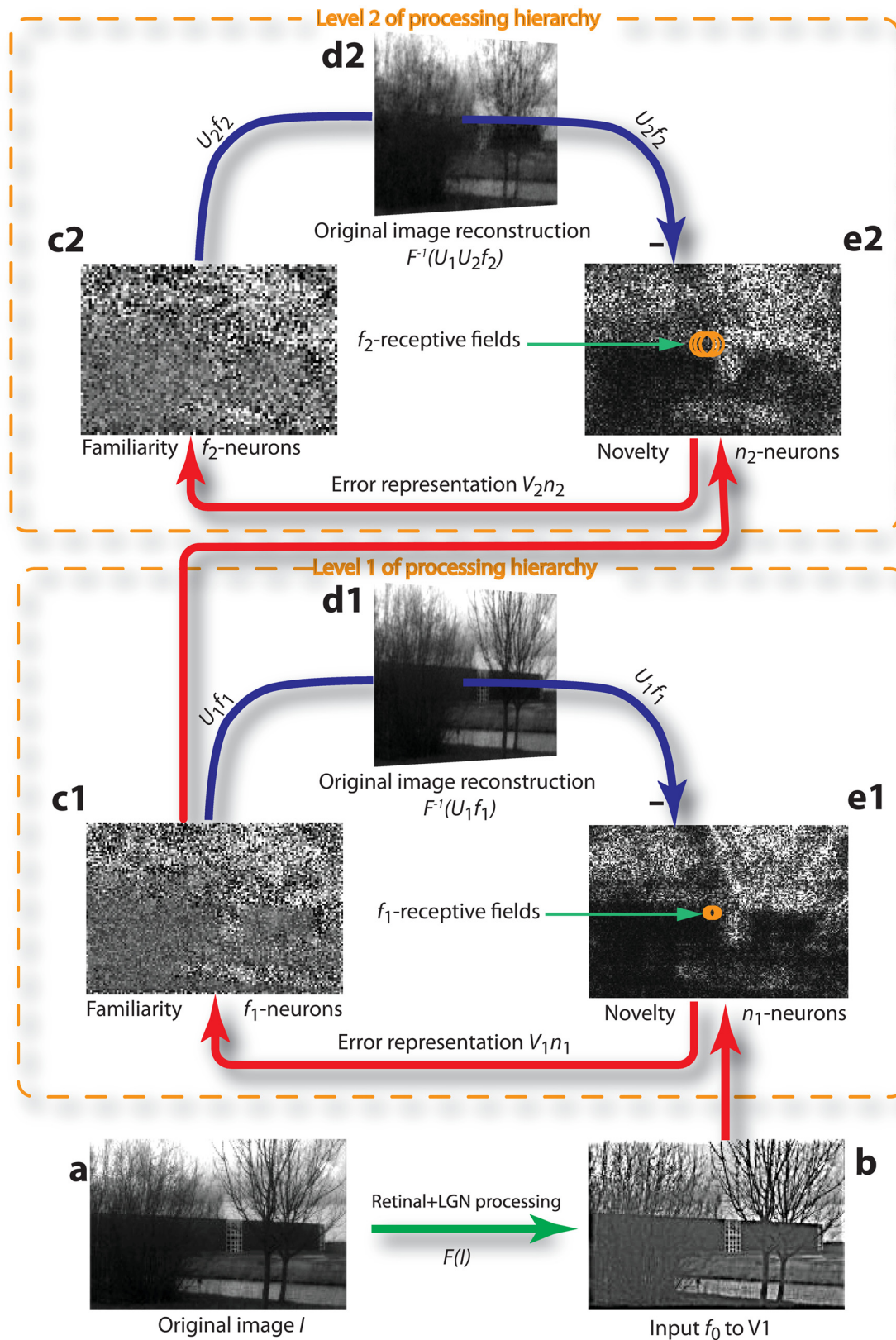
A major challenge in understanding the human connectome is to unravel the intimate relationship between anatomical and effective (functional) connectivity [1–4]. It has been recognized that effective connectivity in terms of correlated activity does not necessarily require direct anatomical projections [5]. On the other hand, anatomical connectivity in the form of white matter tracts has been found to imply effective connectivity [6]. However, as we point out here, this does not hold for anatomical projections in general. Even if excitatory connections

instantaneously drive the postsynaptic activity, in a recurrent network the activity on the time scale of network dynamics may be causally unrelated to the instantaneous drive.

Our model of stimulus representation in the visual system provides an example of how effective and anatomical connectivities may differ. The visual system has been described as a hierarchy of predictive coding schemes, where activity in a lower level is ‘predicted’ by activity in the next level of the hierarchy [7–9]. Because the ‘prediction’ in the higher level is effectively subtracted from the lower-level activity, it was argued that in this lower level only the error signal remains, and, as a consequence, only novelty information is passed to higher cortical levels [7, 10]. However, this conclusion appears to rely on the short-term network dynamics only. Instead, we propose that once the recurrent network has equilibrated, it is in fact the familiarity information, not the novelty, that is projected forward to higher cortical areas (see also [11]). We show that at each level of the hierarchy the recurrent dynamics divides the input signal into orthogonal familiarity and novelty components. Both of these components do effectively only depend on the lower-level familiarity component, but this functional dependence is not reflected in the direct anatomical connectivity (cf. Figs 1 and 2). This is where our approach differs crucially from the previous hierarchical predictive-coding work [7, 11], in which explicit feedback connections from higher cortical areas played an essential role in the network functionality.

To link anatomical and effective connectivity one may start with either of them. Classically, data on anatomical connectivity is first collected and then interpreted in functional terms. From a theoretical point of view, however, it is natural to first consider a possible functional purpose, and then seek its neuronal implementation. Generative models represent a general framework for operating in this opposite direction [12, 13]. The basic idea is to specify a model describing how sensory stimuli can be reconstructed (‘generated’) from a lower-dimensional neuronal activity pattern. The function assigned to the visual system in this setting is to represent visual stimuli in a compressed form such that the original stimuli can be reconstructed as closely as possible (see also [14, 15]). This approach is also adopted in predictive coding [7] and can itself be deduced from a unifying Bayesian optimization principle [16]. Given a generative model for reconstructing an image  $I$  from neuronal firing rates  $f$ ,  $I \approx \Phi(f)$ , one may ask whether the firing rates in turn could be explained by some neuronal processing triggered by the image. This amounts to inverting the generative model and obtaining the firing rates from the image,  $f \approx \Phi^{-1}(I)$ . The task is then to find a generative function  $\Phi$  such that its inverse  $\Phi^{-1}$  can be implemented in a neuronal circuitry. We show that the requirement of neuronal implementability of the inverse generative function strongly constrains the neuronal transfer function, essentially only allowing threshold-linear neurons. Further restrictions on the generative model arise from the fact that the receptive fields of neurons have limited size; thus, additional neuronal layers are required to extract more global features from visual stimuli.

Here we suggest that stimulus representation and recognition in the visual system is based on a modular hierarchy of predictive coding schemes with an effectively feedforward character. Recurrent connections are restricted to each individual level to separate novelty from familiarity information; they do not feed back to preceding levels as originally suggested [7]. We show that a quadratic regularity constraint can make this modular architecture functionally very similar to the fully recurrent architecture while keeping the advantages of a level-specific optimal encoding and the fast relaxation time characteristic of visual perception [17–19]. Furthermore, non-classical receptive field (RF) properties—as observed in the original predictive coding scheme [7, 20]—emerge from the within-level lateral connections, despite the absence of the top-down projections and despite linear neuronal dynamics only. This becomes possible because the RFs are limited to small overlapping areas on which the neurons develop specific interactions.



**Fig 1. Modular predictive coding and image reconstruction after learning.** **a** Example of a novel image  $I$  presented to the network after learning on 1000 other images in a total of 5000 randomly sampled presentations. **b** Input  $f_0$  to the first level after retinal and LGN processing. **c1** Activity of level-1 familiarity

neurons  $f_1$  receiving 1-to-1 input from  $n_1$ -neurons through constrained connection matrix  $V_1 (\approx U_1^T)$ . **d1** Reconstruction of the preprocessed image based on the steady-state activity of the level-1 familiarity neurons,  $f_0 \approx U_1 f_1$ . **e1** Activity of level-1 novelty neurons  $n_1$  receiving 1-to-1 input from the LGN and localized input from  $f_1$ -neurons through the constrained connection matrix  $U_1$ . Orange circles represent RFs of three neighboring  $f_1$ -neurons. **c2, d2, e2** The corresponding quantities for level 2. The number of  $f_2$ -neurons is 50% of the number of  $f_1$ -neurons and 25% of the number of  $n_1$ -neurons (= number of pixels in the visual input).

doi:10.1371/journal.pone.0144636.g001

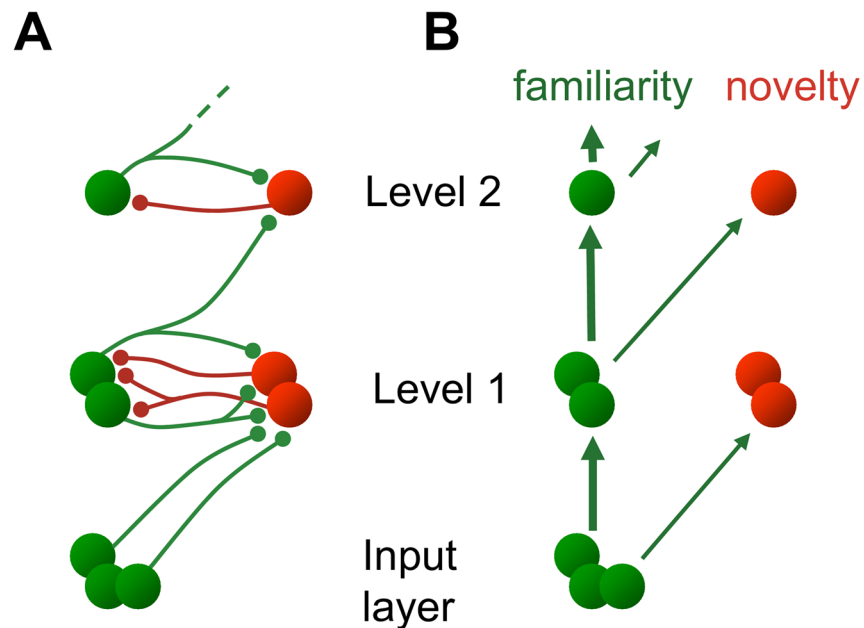
Finally, we suggest that, instead of being essential to the representation of natural visual stimuli by means of predictive coding, top-down connections are engaged in attention, memory recall and in top-down-gated learning, without substantially affecting the fast processing of sensory stimuli.

## Results

### Modular novelty-familiarity coding

In this paper we consider hierarchical coding of visual stimuli where at each level of the hierarchy it is possible to infer ('predict') the activity at the preceding lower level. Mathematically, vector  $f_i$  of neuronal firing rates at level  $i$  can approximate the firing rate vector  $f_{i-1}$  at the lower level,  $f_{i-1} \approx \phi(U_i f_i)$ , where  $U_i$  is a linear transform and  $\phi$  the generative function, a possibly nonlinear function applied component-wise to the linear combinations of neuronal activities at level  $i$ . To enforce information extraction at each level, complexity constraints are imposed on  $f_i$ , e.g., that  $f_i$  be of lower dimension than  $f_{i-1}$ . The approximation quality is measured by a quadratic error,

$$E_i = \frac{1}{2} \| f_{i-1} - \phi(U_i f_i) \|^2, \tag{1}$$



**Fig 2. Anatomical versus effective connectivity.** **A** Schematic anatomical connectivity pattern in the early ventral visual cortex shows recurrent synaptic connections within each level (Eq (4)). Lower-level 'familiarity' neurons (green) project to 'novelty' neurons (red) at the next higher level. **B** Effective connectivity expressing causal relationships results in a purely feedforward network (Eq (5)). At each level, familiarity and novelty information is extracted from the familiarity representation at the previous level.

doi:10.1371/journal.pone.0144636.g002

where for the first level ( $i = 1$ ) the input activity represents the image,  $f_0 = I$ . Within the classical predictive coding framework [7], the total error function  $E_1 + E_2 + \dots + E_L$  is minimized across  $L$  levels, with the consequence that the activity vector  $f_i$  depends on both the activities of the lower and higher levels.

Here, we suggest a modular hierarchical coding which assumes that at each level  $i$  the corresponding error function  $E_i$  is minimized independently of the representation at the higher level (Fig 1). The minimization is achieved both on the time scale of the fast neuronal dynamics and that of the slow synaptic plasticity. For the neuronal dynamics this amounts to calculating the gradient of  $E_i$  with respect to the neuronal firing rates  $f_i$  at each level separately, and equating the negative of this gradient to the temporal derivative of  $f_i$ . Assuming that  $\phi$  is the identity function,  $f_i$  then evolves as

$$\tau \dot{f}_i = -\frac{\partial E_i}{\partial f_i} = U_i^T (f_{i-1} - U_i f_i), \tag{2}$$

with some time constant  $\tau$  and  $U_i^T$  being the transpose of matrix  $U_i$ . As we argue below, nonlinear functions  $\phi$  different from threshold-linear would require non-local neuronal processing rendering the corresponding generative model biologically unlikely, at least in the absence of non-monotonic gain modulation.

To make the dynamics Eq (2) neuronally plausible we introduce auxiliary ‘novelty’ neurons which represent the difference  $n_i = f_{i-1} - U_i f_i$ . Note that this difference expresses a prediction error, i.e., the residual activity in the lower-level neurons  $f_{i-1}$  that cannot be ‘predicted’ by the higher-level neuronal activities  $f_i$ . In our interpretation, this difference is calculated by recurrent connections within the upper layer  $i$ , without assuming top-down connections (Fig 2a). Since in reality and in our model the reconstruction is learned based on repeated stimulus presentation, the  $f_i$  neurons encode the lower-level activity by exploiting the statistics of all images presented, and hence we refer to the  $f_i$ ’s as ‘familiarity’ neurons (also called ‘prediction’ or ‘representation’ neurons—cf. [21]). Since due to the above definition the activity of novelty neurons tracks the prediction errors instantaneously, their neuronal time constant must be short compared to that of the  $f_i$  neurons. This dynamical constraint can be taken into account by introducing a small leak term  $-\epsilon f_i$ , yielding a long integration time constant for  $f_i$  as compared to the dynamics of  $n_i$ . The leak term can also be considered as additional constraint that keeps the overall activity of the  $f_i$  neurons low,

$$E_i = \frac{1}{2} \| I - U_i f_i \|^2 + \frac{\epsilon}{2} \sum_{k=1}^{N_i} (f_i)_k^2. \tag{3}$$

For biological realism—and to allow for nonlinear computational properties (see [22–24] and below)—we truncate the firing rates of the  $f_i$  neurons at 0 whenever they would become negative otherwise. The overall neuronal dynamics minimizing the individual error functions  $E_i$  then becomes

$$\begin{aligned} \tau_f \dot{f}_i &= -\epsilon f_i + V_i n_i, \quad \text{constrained to } f_i \geq 0 \\ \tau_n \dot{n}_i &= -n_i + f_{i-1} - U_i f_i, \end{aligned} \tag{4}$$

where  $U_i$  represents the matrix of synaptic weights from the  $f_i$  to the  $n_i$  neurons, and its approximate transpose  $V_i \approx U_i^T$  the weights from the  $n_i$  to the  $f_i$  neurons within level  $i$  (similar, but not the same initial values—see Methods—are more realistic biologically and the transposed update rule Eq (6) lead to  $V_i$  approaching  $U_i^T$  as learning progresses). Similar neuronal

dynamics applied to a single layer with the whole image as each neuron’s receptive field was introduced in [22].

The same quadratic constraint in Eq (3) also mimics the effect of the missing top-down connections that would introduce a quadratic penalty term on the components not represented by the upper level (Eq (S.7) in S1 Supporting Information). In the cross-level predictive coding scheme, the top-down connections would selectively suppress components that provide less information for the coarse-grained representation at the higher level. But in doing so, the lower-level network will only converge to a steady state when the higher-level network does. This deteriorates the convergence time for the lower level towards that of the higher level, which itself can only extract the relevant information when the lower level dynamics is near relaxation. The dynamics Eq (4), instead, is faster as it does not depend on the more global  $f_{i+1}$  activity.

### Anatomical versus effective connectivity

The dynamics in Eq (4) describes a layered hierarchy of mutually connected familiarity and novelty neurons  $f_i$  and  $n_i$ , respectively, which could be embedded in the visual system. Starting with an image  $f_0 = \tilde{I}$  preprocessed by the lateral geniculate nucleus (LGN), novelty neurons  $n_i$  receive feedforward input from familiarity neurons  $f_{i-1}$  of the lower level as well as input from familiarity neurons  $f_i$  of the same level. These latter also receive input from novelty neurons  $n_i$  of the same level, and are thus embedded in a recurrent network within level  $i$  (Fig 2a). However, in the steady state we can express the activity of both the familiarity and novelty neurons as a function of input from the previous level: from Eq (2) with  $\dot{f}_i = 0$  we obtain  $f_i = U_i^+ f_{i-1}$ , where  $U_i^+$  is the pseudoinverse of  $U_i$ . Plugging this into the second equation of Eq (4) while setting  $\dot{n}_i = 0$  yields  $n_i$  as a function of  $f_{i-1}$ . Hence, while the recurrent anatomical connectivity is expressed by Eq (4) (Fig 2a), the effective connectivity in the steady state becomes purely feed-forward (Fig 2b):

$$f_i = U_i^+ f_{i-1} \quad \text{and} \quad n_i = (1 - U_i U_i^+) f_{i-1} \tag{5}$$

This effective connectivity represents the underlying causalities and effectively drives the responses of familiarity and novelty neurons at level  $i$  by those of familiarity neurons at level  $i - 1$ , averaged across the time scale  $\tau_f$  of the network dynamics (on the order of milliseconds). Due to the modularity, the full representational power at a specific level  $i$  is exploited to optimize the novelty-familiarity representation at that spatial resolution, while also enabling an optimized, parallel readout from the different levels. In contrast, when optimizing only the single sum  $E_1 + E_2 + \dots + E_L$  across all levels as in the classical ‘cross-level’ predictive coding [7], top-down connections from level  $i+1$  introduce additional constraints at level  $i$  and help to reduce the error at level  $i+1$ , but this is at the expense of a non-optimal lower-level representation (see Eq (S.7) and Fig A in S1 Supporting Information). A compromise that helps to improve the prediction error at the next level  $i+1$  while retaining the full representational power at level  $i$  is to introduce a quadratic penalty term on the familiarity neurons at level  $i$  that we used to derive the dynamics for  $f_i$  (Eqs (3) and (4)).

### Synaptic plasticity and hierarchical PCA

Learning was implemented in our model by further minimizing each individual error function  $E_i$  with respect to the synaptic connectivity matrix  $U_i$  on the slow time scale of stimulus presentation. While the input to the first level is clamped at the currently presented image  $f_0 = \tilde{I}$ , the neuronal dynamics is relaxed across the levels until it reaches a steady state, and  $U_i$  is then

updated by gradient descent on the error function  $E_i$  (Eq (1)) with respect to synaptic weight parameters  $U_i$ . Subject to a spatial locality constraint ensuring that outside a small area of the RF the connectivity is always 0, the weight update is

$$\Delta U_i = -\eta \frac{\partial E_i}{\partial U_i} = \eta (f_{i-1} - U_i f_i^*) (f_i^*)^T = \eta n_i^* (f_i^*)^T \quad (6)$$

with some learning rate  $\eta$ . Here  $f_i^*$  and  $n_i^*$  are steady-state neuronal activities after network relaxation. Notice that this synaptic update rule has the Hebbian form of postsynaptic activity times presynaptic activity. Similarly, the synaptic updates from the novelty to the familiarity neurons take the Hebbian form

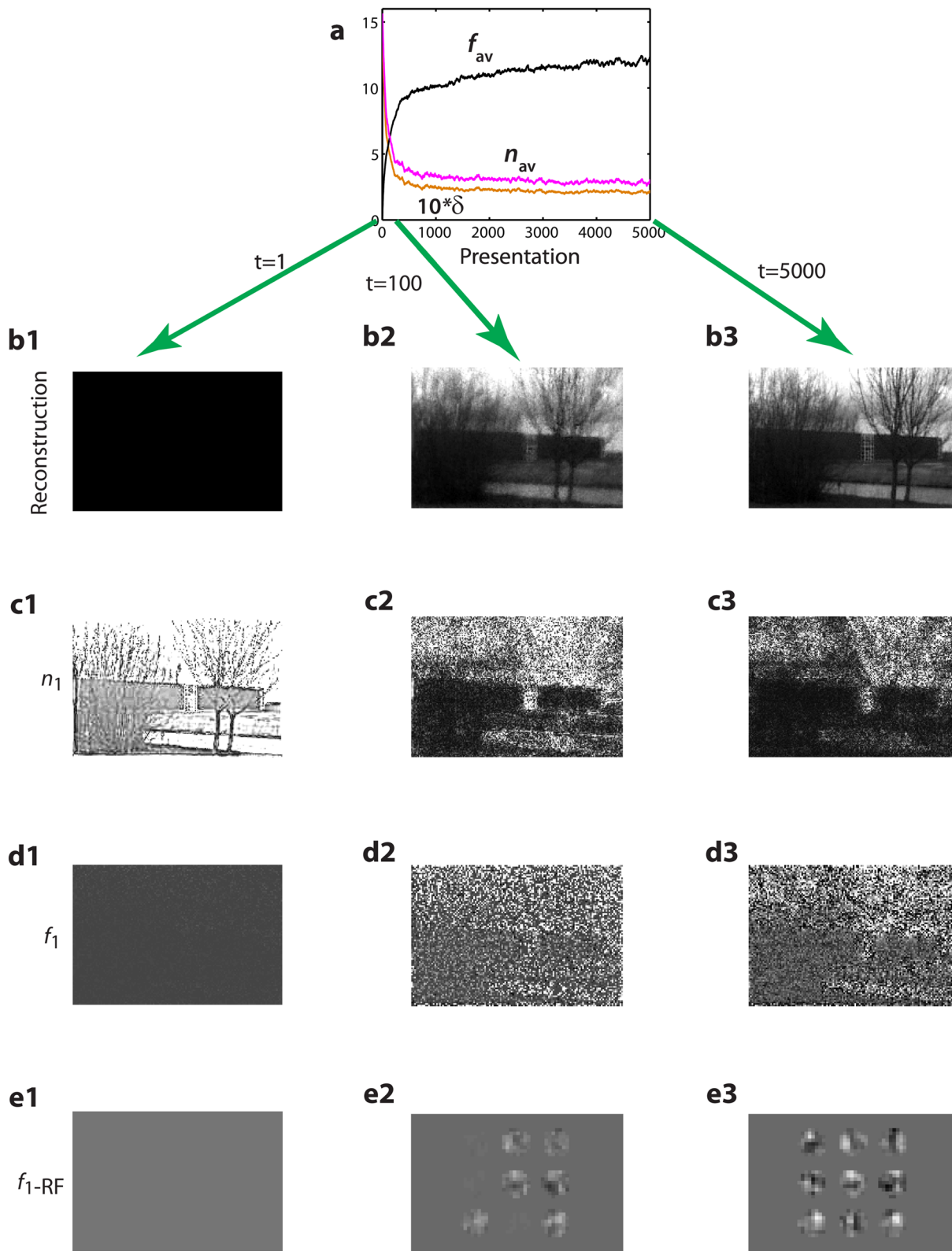
$$\Delta V_i = \eta f_i (n_i)^T \quad (7)$$

In the present case of modular hierarchical coding (but not for the cross-level predictive coding [7]), this architecture performs a hierarchical principal-component analysis (PCA) of the image. PCA is known to minimize the mean squared error ([25, 26]; also demonstrated explicitly in [S1 Supporting Information](#), Section S.II). The familiarity neurons' activities  $f_i$  represent the principal components of the activities  $f_{i-1}$  of the lower-level familiarity neurons by virtue of the effective feedforward connectivity (Eq (5)). But unlike in the previous work [22, 25, 26], the receptive field (RF) of a single  $f_i$  neuron is limited to a small area and does not span the whole image. Different components are extracted on the RF overlaps. These components jointly span the space of the first principal modes on the overlaps (cf. [Fig 3e](#) and [S1 Supporting Information](#), Subsection S.II). Yet, the PCA property does not imply that the image analysis is linear. Due to the thresholded transfer functions, neurons with overlapping or neighboring RFs may nonlinearly interact to minimize the prediction error on the entire set of images (see Subsection on endstopping below).

### Development of level-1 recurrent connectivity

To track the progress of learning, we considered the evolution of the reconstruction error on a training set of 1000 natural images and the average steady-state activities at the first level. As expected, the average activity  $f_{av}$  across the familiarity neurons  $f_1$  increases while the average activity  $n_{av}$  across the novelty neurons  $n_1$  decreases with repeated presentations of the images ([Fig 3a](#), black and magenta curves, respectively). At the same time, the reconstruction of the LGN-preprocessed image  $f_0$  based on the activity of familiarity neurons  $f_1$ ,  $f_0 \approx U_1 f_1$ , becomes more accurate as expressed by the error curve ([Fig 3a](#), golden) and the example reconstructions ([Fig 3b](#)). The topographic representation of the neuronal activities shows that during the learning process the contrast among the novelty neurons decreases ([Fig 3c1–3c3](#)) while among the familiarity neurons it increases ([Fig 3d1–3d3](#)). Learning transforms novelty into familiarity while keeping the original information (as expressed by  $f_0 = U_1 f_1 + n_1$ ).

Inspection of the receptive fields (RFs) of familiarity neurons  $f_1$ , as expressed by the vector of synaptic input strengths from  $n_1$  neurons (rows of  $V_1$ ), shows RFs composed of patches of excitation and inhibition ([Fig 3e](#)). When combined to jointly cover the input space, they form the first principal components of the correlation matrix of the inputs ([S1 Supporting Information](#), Subsection S.II). An additional sparseness constraint in the energy functional (Eq (1)), e.g., an additional penalty term for the norm of  $f_i$  or  $U_i$ , see [S1 Supporting Information](#), may force the weight vectors to become orthogonal, while the RFs become more Gabor-like, similar to the ones observed biologically [27–29]. Yet, as the characterization of the RFs depends on the choice of stimuli [30–32], we did not intend to reproduce specific RF shapes.



**Fig 3. Local recurrent connectivity emerging from unsupervised learning improves reconstruction quality.** **a** Evolution of reconstruction error ( $\delta$ , golden, averaged across 50 consecutive presentations) during 5000 random presentations from a set of 1000 training images. Average activity of level-1 novelty neurons ( $n_{av}$ , magenta) mirrors decrease in  $\delta$ . Initial sharp decrease in  $n_1$  activity is explained by average activity increase of level-1 familiarity



neurons ( $f_{av}$ , black), which quickly learn to extract the most dominant component, mean local brightness (cf. d2). **b1–b3** Reconstruction of a single novel image (not used for training) after 0, 100, and 5000 presentations based on  $f_1$  activities. **c1–c3** Activity of all novelty neurons  $n_1$  (same number of neurons as pixels in image), showing the reduction of local novelty with decreasing reconstruction error. **d1–d3** Corresponding activities of  $f_1$ -neurons in response to the novel image. Image representation is gradually refined, despite the image not having been presented to the network. **e1–e3** Evolution of overlapping receptive fields (RFs, rows of  $V_1$ ), shown separated for visualization, of nine representative nearest-neighbor familiarity neurons.

doi:10.1371/journal.pone.0144636.g003

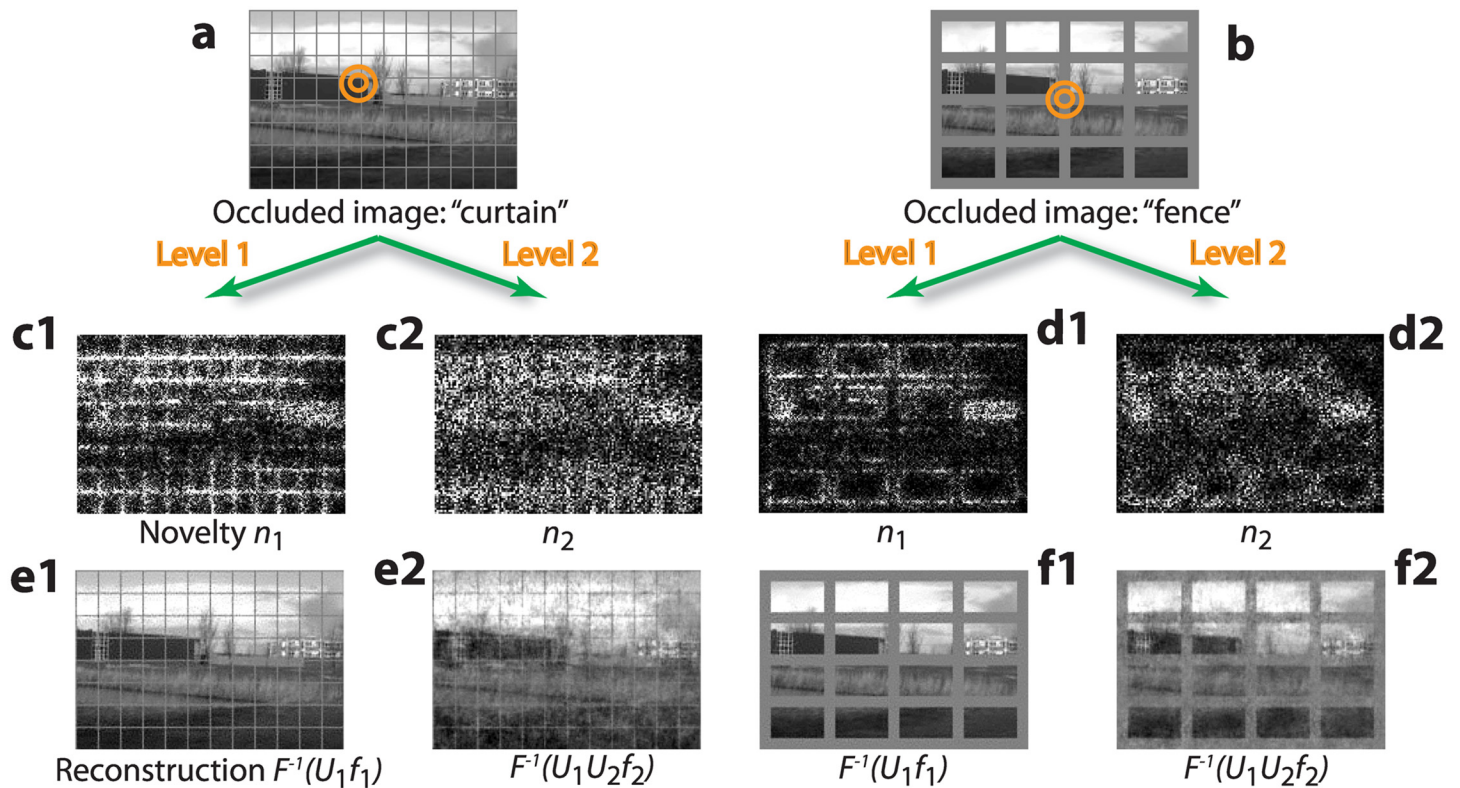
After learning, we assessed the learning quality on a set of 200 novel images from the same library but never presented to the network before. The generalization quality is very good: the average reconstruction error was less than 5% larger relative to that for the images used in learning. Note that a specific image was only presented approximately 5 times (in each of the 5000 presentations, an image was randomly selected out of the 1000 images) during the learning process, and so learning of the local image structure is based on the statistics of all the 1000 images.

## Hierarchical novelty-familiarity representation

To illustrate how the image representation is decomposed into familiarity and novelty signals at different spatial scales, we superimposed a grid of two different line widths onto an image not used for learning (Fig 4a and 4b). Neither grid is a typical feature of natural images, and hence they are detected as novel. However, because of the different line widths, the narrow ‘curtain’ is mainly detected at level 1 (white grid in Fig 4c1, reflecting activity of  $n_1$ -neurons), while the area of the wide ‘fence’ is mainly detected at level 2 (white grid in Fig 4d2, reflecting activity of  $n_2$ -neurons). The wide bars of the fence are recognized as statistically familiar at level 1 because a level-1 RF is covered by a uniform bar of the fence (narrow circle in Fig 4b), and the uniform brightness as a zero-order principal component can be easily reconstructed by the familiarity neurons. At this first level, only edges along the bars of the fence are partially detected by the novelty neurons (Fig 4d1).

Novelty is a local phenomenon restricted to the receptive field. Where novelty is detected, the original image can be partially reconstructed using surrounding familiarity neurons through lateral connections. In fact, image reconstruction based on the familiarity neurons alone partially succeeds to retouch away the narrow curtain at level 1 (Fig 4e1), and both the fence and the curtain at level 2 (Fig 4e2 and 4f2). As the accurate image reconstructions from the fewer level-2 familiarity neurons show, familiarity information is still available at that level, despite the fact that input to level 2 is only fed into level-2 novelty neurons. This is possible because familiarity neurons (at level 2) continuously integrate incoming information from novelty neurons, until the activity in the novelty neurons cannot be explained anymore by the familiarity neurons (cf. Eq (4)).

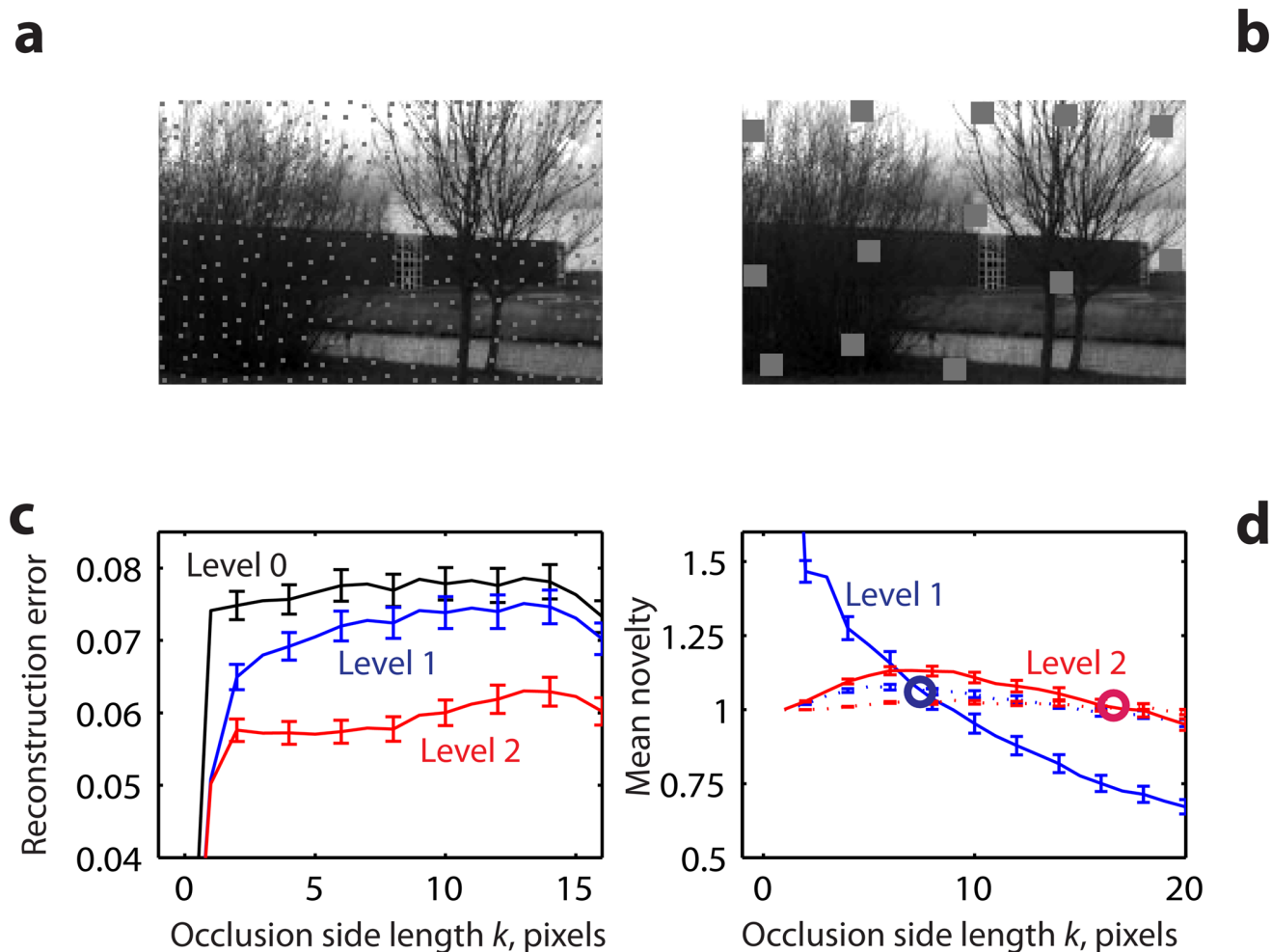
The size of an occlusion that can be correctly reconstructed depends on how small it is compared to the RFs of the corresponding familiarity neurons. To investigate how the reconstruction quality at the first two levels improves as a function of the occlusion size, we consider randomly scattered square-shaped occlusions of the same total area (5% of the image area, Fig 5a and 5b). Reconstruction quality is assessed based on 200 novel images from the same library, but again not part of the training set. As expected, level-1 and level-2 familiarity neurons carry original image information even for occluded patches. The distance between the reconstructions of occluded and non-occluded images is much less than that between the non-occluded and occluded images themselves for small occlusion sizes ( $k = 0$  corresponds to no occlusion and, correspondingly, all curves start at 0 for  $k = 0$ ). This suggests that for small occlusion sizes despite the occlusions, it is the original, non-occluded, image that is reconstructed.



**Fig 4. Novelty detection and familiarity information fill-in at different spatial scales.** A familiar image is either occluded by a narrow, 1-pixel-wide ‘curtain’ (a) or by a 7-pixel-wide ‘fence’ (b). Circles represent the RFs of level-1 (smaller circle) and level-2 neurons, respectively. Novelty neurons at level 1 are activated by the fine curtain (c1), while at level 2, curtain information is filtered out (c2). For the fence, novelty at level 1 is only detected at some edges (d1), but novelty of the wide fence bars is detected at level 2 (d2). Reconstruction of the image (LGN activity) based on the level-1 and level-2 familiarity neurons for the curtain (e1, 2) and the fence occlusions (f1, 2) shows how the occluded parts of the image are filled in despite the reduced number of neurons.

doi:10.1371/journal.pone.0144636.g004

The activities of the novelty neurons on the occluded and non-occluded areas represent a biologically feasible measure of reconstruction errors in the respective image areas. Level-1 novelty neurons’ mean activity over the occluded areas (Fig 5d, decreasing solid blue curve) is highest for the smallest occlusion patch size  $k$  and diminishes progressively with increasing  $k$ . It equals the non-occluded neurons’ activity (Fig 5d, dotted blue curve) at approximately  $k = 8$  pixels (the diameter of a single level-1 RF), and then continues to decrease, indicating that uniform occlusions larger than the size of a single RF are indeed the most familiar and easiest to reconstruct. For level-2 novelty cells, the occluded (Fig 5d, solid red curve) and non-occluded (dotted red curve) mean activities become equal at approximately the size of the level-2 RF (diameter  $\approx 18$ ), as expected. The solid red curve, unlike the blue one, begins at 1 for the smallest occlusion of size  $k = 1$  (and remains non-monotone for small occlusion sizes). This is a consequence of the natural definition of the RFs of level-2 novelty neurons as equal to those of the corresponding level-1 familiarity neuron into which the occlusions fall. E.g., for  $k = 1$  the 5% total occlusion translates into approximately 1 pixel of occlusion for every 4.5 pixels of the image and hence, on average, every level-2 novelty neuron’s RF contains a pixel of the occlusion when  $k = 1$ .



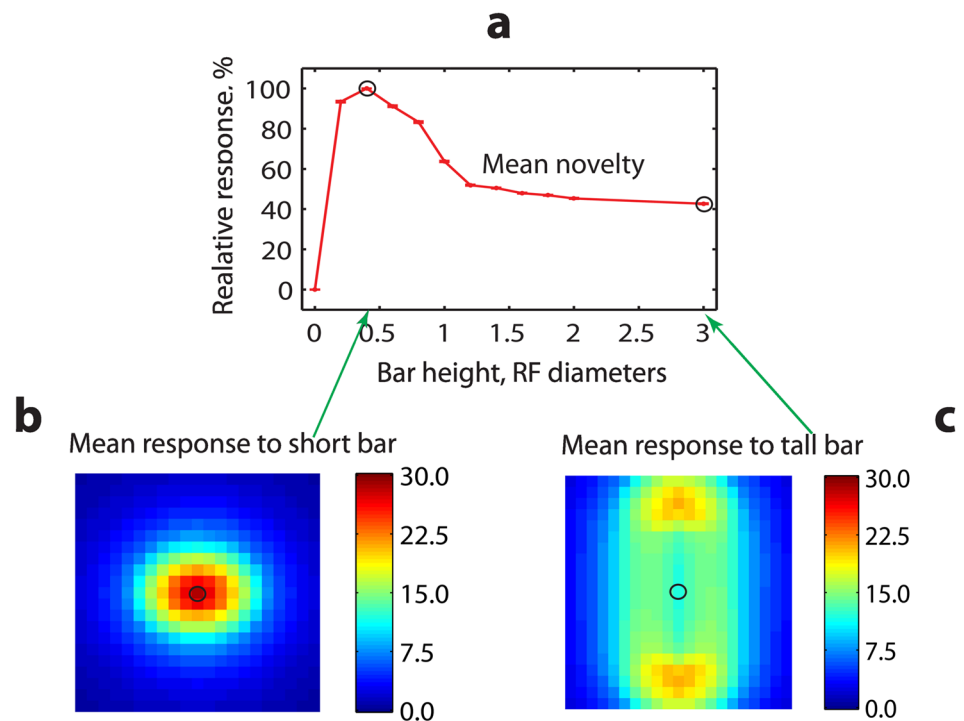
**Fig 5. Reconstruction and novelty as a function of occlusion patch size  $k$ .** **a, b** Example of a novel image with occlusion patches (grey squares) of side length  $k = 1$  (**a**) and 10 pixels (**b**). For each patch size, the total occluded area is 5% of the image area. **c** Reconstructions of occluded and non-occluded images progressively become more similar across levels. Average distance  $\delta$  (across 200 novel images) between the non-occluded original image  $I$  and the same image  $\hat{I}$  occluded by the patches (black curve, 'level 0'), and between the reconstructions of those images  $F^{-1}(I)$  and  $F^{-1}(\hat{I})$  based on the familiarity neurons at level 1 (blue) and level 2 (red), respectively. **d** Average activity of level-1 (blue) and level-2 (red) novelty cells over the occluded (solid) and non-occluded image areas (dotted curves). Novelty neurons in the occluded parts of the image are more active than novelty neurons in the non-occluded parts of the image as long as the patch size  $k$  is smaller than the RF diameter (8 and 18 pixels for levels 1 and 2—cf. blue and red circles—respectively).

doi:10.1371/journal.pone.0144636.g005

### Non-classical RF property of endstopping results from within-layer connectivity

To test for the non-classical RF property of endstopping, we stimulated each level-1 familiarity neuron separately by a uniform bar extending horizontally across the whole neuron's RF and vertically across 40% of the RF diameter. The bar was approximately 5 times brighter (similarly to [7]) than the background, whose brightness was equal to the average over all training images. The equilibrium activity of each level-1 familiarity cell's response to a horizontal bar was then recorded and the 20% of the neurons with the highest equilibrium activities were designated bar detectors. These detectors were then stimulated, again one at a time, by bars of the same width (equal to the RF diameter) and height varying systematically from 0 to 3 RF diameters. The activity deviations of all the novelty cells within each bar detector's RF from their rest values, averaged over each bar detector's RF, were computed for each height.

The relative response of the novelty neurons decreased when the bar height extended beyond 0.4 RF diameter (Fig 6a) indicating cooperative effects between the familiarity cells with overlapping RFs. To illustrate the typical behavior of novelty neurons, we also averaged novelty activity across RFs of the familiarity neurons in a fixed-size neighborhood of approximately 3.5-by-3.5 RF diameters around each bar detector. The responses to the short stimulus of height 0.4 RF diameter (Fig 6b) and the tallest stimulus of height 3 RF diameters (Fig 6c) show that for the tall stimulus, despite the same brightness, the activity is considerably lower and more uniform. The reduction of novelty activity in the bar center with increasing bar height reveals an effective polysynaptic inhibitory effect of the neighboring neurons on the bar detectors, consistent with the endstopping behavior of bar-detecting neurons in the primary visual cortex arising from lateral connectivity [33–35]. The comparison between the response to the short bar (Fig 6b) and that to the tall bar (that itself can be considered as a juxtaposition of 3 vertically aligned bars of height 1) reveals the nonlinear inhibitory processing due to the thresholding of the  $f_i$  activities. In fact, the response to the tall bar (Fig 6c) is much weaker than the sum of the responses to the bars of height 1 would be.



**Fig 6. Non-classical effect of endstopping is present in the effectively feedforward architecture of modular predictive coding.** **a** Activity of level-1 novelty neurons averaged across the RF of bar-detecting level-1 familiarity neurons decreases as bar height exceeds approximately one-half of the RF diameter. Activity is shown relative to baseline activity for the short bar of 0.4 RF diameter and five times brighter than the background. Black circles indicate the responses of bar detectors relative to the baseline for short and tall bars in panels b and c, respectively. **b** Average activities of level-1 novelty neurons in a neighborhood of 3.5-by-3.5 RF diameters around each bar-detector in response to the short bar. Central dark red square (black circle) represents the average activity of all novelty cells within the RF of a bar detector, further averaged across all bar detectors. The same double averaging was applied to the novelty neurons within the RFs of the neighboring familiarity neurons. While these RFs are highly overlapping, they are displayed here side-by-side. **c** Same as in panel b, but in response to a tall bar of height 3 (instead of 0.4) RF diameters. The averaged activity of novelty neurons (representing the local reconstruction error) within the bar-detectors RF (black circle) is lower than that for the short bar (compare black circles) and for the same reason also lower than that recorded at the ends of the bar (top and bottom yellow patches).

doi:10.1371/journal.pone.0144636.g006

## Nonlinear predictive coding is neuronally implementable for threshold-linear transfer functions only

Predictive coding has been considered within a general framework of optimization principles allowing for nonlinear feature extraction [16, 21, 36]. In striving for a neuronal implementation of such nonlinearities, however, we find below that it is essentially only the linear neuron model with the quadratic error function that can be implemented by neurons using locally available information. Yet, threshold-linear neurons allow the predictive coding model to remain linear while enabling nonlinear feature extraction. Given the representation of the unconstrained activities  $f$  by a difference of threshold-linear ‘ON’ and ‘OFF’ neurons,  $f = [f] - [-f]$ , another readout neuron may easily extract the sum  $[f] + [-f]$ . Qualitatively, the latter operation is similar to taking the square of the linear filter output,  $f^2$ , which was shown to lead to phase-invariant receptive fields of complex [37] or motion-selective cells [38]. Hence, the neuronal implementation of the linear version of modular predictive coding yields the ingredients to explain the ON-OFF simple cells [39] and also complex cells in the primary visual cortex (V1) as they arise in nonlinear optimization models [40, 41].

**Implementation of positive transfer functions.** As everywhere in this paper, the neuronal activities  $f_i$  were kept positive by truncating them at 0 should they become negative. We also explored the option of a non-negative prediction error. To achieve this, we applied a shift to the  $n$ -activities by a vector  $n_o$  with equal and positive components (large enough so that the  $n_i$  became nonnegative at all times) and then subtracted this shift again to calculate  $\hat{f}_i$  and the weight changes as if  $n_i$  had not been offset. Hence, instead of Eqs (4), (6) and (7) we consider

$$\begin{aligned} \tau_f \dot{f}_i &= -\epsilon f_i + V_i n_i - b_i, \text{ constrained to } f_i \geq 0 \\ \tau_n \dot{n}_i &= -n_i + n_o + f_{i-1} - U_i f_i \end{aligned} \tag{8}$$

$$\begin{aligned} \Delta U_i &= \eta (n_i^* - n_o) (f_i^*)^T \\ \Delta V_i &= \eta f_i^* (n_i^* - n_o)^T \\ \Delta b_i &= \eta^b f_i^* \end{aligned} \tag{9}$$

Here, again,  $f_i^*$  and  $n_i^*$  are the steady-state neuronal activities after network relaxation, while  $\eta$  and  $\eta^b$  represent learning rates. The bias  $b_i$  must converge towards  $V_i n_o$  due to the steady state conditions  $\langle n_i \rangle = n_o$ , which express that the prediction errors  $n_i - n_o$  are balanced around 0. This is the case because  $f_i^*$  on the right-hand side of the  $b_i$ -update equation is a locally available approximation to the negative gradient of  $(V_i n_i^* - b_i)^2$  with respect to  $b_i$ .

**Nonlinear generative functions do not have a neuronal implementation.** When considering a nonlinear generative function  $\phi$ , the gradients of Eq (1) with respect to  $f_i$  and  $U_i$  become, instead of Eqs (2) and (6),

$$\tau_f \dot{f}_i = U_i^T ((f_{i-1} - \phi(U_i f_i)) \cdot * \phi'(U_i f_i)) \tag{10}$$

$$\Delta U_i = \eta ((f_{i-1}^* - \phi(U_i f_i^*)) \cdot * \phi'(U_i f_i^*)) (f_i^*)^T, \tag{11}$$

where  $\cdot *$  is the componentwise multiplication,  $\eta$  is a positive learning rate and  $f_i^*$  is the steady-state activity.

Any nonlinear generative function  $\phi$  that is different from threshold-linear introduces a nontrivial multiplicative modulation of the synaptic input to the  $f_i$ -neurons as expressed by the

pointwise product in Eq (10). The same product also arises in updating the weight matrix  $U_i$  in Eq (11). By again considering the difference in the above equations as auxiliary neuronal quantities,  $n_i = f_{i-1} - \phi(U_i f_i)$ , the modulation by  $\phi'(U_i f_i)$  even becomes non-local. In fact, the input  $f_i$  weighted by the synaptic strengths,  $U_i f_i$ , would then become the input to a  $n_i$  neuron and so it is not locally available at the site of the  $f_i$  neuron (without assuming a specific rewiring and duplication of synaptic weights).

Alternatively, one may postulate a neuron with highly non-monotonic interactions among different synaptic inputs as expressed by the steady state equation of that neuron,  $n_i = (f_{i-1} - \phi(U_i f_i)) \cdot \phi'(U_i f_i)$ . This would imply, on the one hand, that some part of the synaptic current  $U_i f_i$  is nonlinearly added to the other input  $f_{i-1}$ , while on the other hand, the total postsynaptic current is multiplicatively modulated by the derivative  $\phi'(U_i f_i)$ . Although multiplicative gain modulation, say by some dendritic input, is possible [42], this modulation would be non-monotonic since  $\phi'$  (for a sigmoidal function  $\phi$ ) is 0 for both small and high values.

**A possible non-quadratic error function.** Further investigating other possible nonlinearities in the optimization problem at hand, we also considered the error function

$$E_{U_i}(f_i, f_{i-1}) = \Psi(f_{i-1} - U_i f_i)$$

with  $\Psi(x) = \frac{1}{\alpha^2} \log \cosh \alpha x$  applied component-wise. For small  $x$  this error is quadratic,  $\Psi(x) \approx \frac{1}{2} x^2$ , and for large  $x$  it is linear,  $\Psi(x) \approx \frac{1}{\alpha} |x|$ . We can combine this nonlinearity with the rectification of  $f_i$  and the upwards shift of  $n_i$ . In the neuronal and synaptic dynamics specified by Eq (8) only the second line then changes to

$$\tau_s \dot{n}_i = -n_i + n_o + \Psi'(f_{i-1} - U_i f_i),$$

where  $\Psi'(x) = \frac{1}{\alpha} \tanh \alpha x \approx x$  for small  $x$  is a standard sigmoidal nonlinearity often considered in modeling a saturating neuronal transfer function. In computational terms, the nonlinearity  $\Psi$  tends to alleviate the effects of statistical outliers in input stimuli. Our simulations (with optimized parameter  $\alpha = 4$ , not shown), however, reveal that these benefits are rather humble. Hence, all of the results presented in this paper are for the simpler, more transparent model described by Eqs (8) and (9).

## Discussion

We have reconsidered predictive coding as an organization principle of the visual cortex. While in the original work the anatomical feedforward propagation of information from novelty neurons has been emphasized [7, 10, 11], we show that in functional terms it is actually the familiarity, not the novelty, information that is fed to the next level. This apparent conundrum arises because the only anatomical connections from a lower to a higher level project from the lower-level familiarity neurons into the higher-level novelty neurons (Fig 2). However, at the higher level, it is again the familiarity information that is extracted from the feedforward input, although now at a coarser resolution. The network dynamics continuously separates familiarity from the novelty information while simultaneously building up both representations.

The modular predictive coding scheme that we are proposing here assumes that on the fast recognition time scale features are only extracted via level-specific recurrent circuitry, not via top-down projections (Fig 1). This modularity is a compromise between the fully backward-connected original coding scheme [7] and a purely feedforward hierarchy without lateral connectivity [1]. Viewing stimulus representation as being modular without top-down feedback has distinct computational advantages. First, the network computation can be understood as a hierarchical principal component analysis with increasing receptive field sizes. Second, it

reduces relaxation times and hence the time for feature recognition at various spatial resolutions that has been shown to be as fast as 30–100ms [17–19]. Third, at each level the full representational power of all neurons is used to optimally extract the information at the corresponding level of granularity. Correspondingly, stimulus compression is achieved solely by limiting the number of prediction neurons, not by shunting the activity by top-down inputs from the higher level.

While our model does not include feedback projections from higher levels, we do not suggest that they would not be of functional use. The ubiquitous top-down connections [43, 44] may have an important functional role in attention gating [45], in memory retrieval initiated from higher cortical areas [46], or in gating synaptic plasticity by a modulatory input from other areas [47–49]. However, we do propose that such connections may not be essential in shaping the stimulus representations on the short-term time scale of the neuronal dynamics. In fact, assuming that the activity of the novelty and familiarity neurons can be read out from all levels of the hierarchy, additional top-down projections from within the predictive coding network cannot provide new information about the stimulus. Learning, on the other hand, provides a helpful means to extract the relevant information by separating novelty from familiarity. While learning strengthens familiarity and makes novelty more salient, the full information remains accessible in the combined novelty-familiarity representation. When restricted to the familiarity neurons, however, the original image is re-represented at each new level in a more compressed form by filling in the within-level predictions of progressively increasing size (Figs 3–5).

Despite the lack of top-down projections—and different from [20] where additional multiplicative and divisive nonlinearities were introduced—our model explains the emergence of non-classical receptive field properties via lateral connectivity alone. Such results are in line with recent experimental findings showing that RFs of V1 cells cannot be defined without considering their spatiotemporal context [30–32, 50]. In our case, the context sensitivity is facilitated by the fact that we did not impose a strict sparseness constraint, but instead derived a ‘soft’ quadratic penalty term from the full predictive coding scheme. In contrast to the sparseness constraint that enforces zero activity, the soft constraint still allows for low activity that may accumulate and shape the non-classical RF properties. Yet, learning still reduces the overall activity within each level and hence leads to a ‘soft sparseness’. This, in addition, arises because learning decreases the prediction error and hence decreases the activity of the numerous novelty neurons.

Considering possible neuronal implementation of predictive coding, we have arrived at severe restrictions on the generative functions prompted by the requirement of compatibility with the current knowledge on neuronal processing. We found that, in essence, the only type of nonlinearity for generative functions that is neuronally implementable is the threshold-linear function, whereas other nonlinearities would require non-local processing. However, use of the threshold-linear transfer function still results in the neuronal processing becoming nonlinear as shown by the non-classical RF properties (Fig 6b and 6c), making the suggested PCA effectively nonlinear. It has been shown that with such a nonlinearity, PCA can extract independent components [23, 24]. We further suggest that the thresholding of the linear transfer function at zero leads to the emergence of ON-cells and OFF-cells [39] that would combine to produce a fully linear response function or to a response function with complex cell properties [41]. How far modular predictive coding as presented here can explain such nonlinear properties of V1 neurons, however, is yet to be analyzed in detail.

## Methods

### Model inputs, structure and fast dynamics

We applied the modular coding scheme to a set of 1000 grey-level images  $I$  (128-by-192 pixels) from a natural image library [51]. They were passed through localized center-surround filters (Fig 1a and 1b) intended to mimic the combined effects of eye adaptation and LGN processing. From each pixel's brightness in the image we subtracted 80% of the mean brightness, including that of the pixel itself, in a small circular neighborhood of radius 5 around that pixel. We implemented this filtering for each image  $I$  by performing a Fast Fourier Transform (FFT) of the image, multiplying it by the FFT of the filter, and then taking the inverse FFT to obtain the filtered image  $\tilde{I} = F(I)$ .

The output of the LGN, an image  $\tilde{I}$  of the same size as  $I$ , was fed into the network with dynamics defined in Eq (4),  $f_0 = \tilde{I}$  and two further levels  $i = 1, 2$ . The truncation of  $f_i$  at 0 ensured that the activities remained positive. In view of this rectification we doubled the number of the familiarity neurons to maintain the reconstruction quality as compared to unconstrained neurons used, e.g., in [7]. The total number of (threshold-linear) familiarity neurons in the first level thus was 2 times, and in the second level 4 times as small as the number of pixels in the image, leading to a gradually compressed image representation across levels. When the truncation of  $f_i$  at 0 was not applied, we obtained very similar results for the significantly higher compression factors of 4 and 16 in the first and second levels, respectively (data not shown). Each filtered image was presented to the network for  $5\tau_f$  time units so that the fast dynamics could equilibrate. To keep the dynamics of the novelty neurons fast as compared to one of the familiarity neurons, we set the  $n_i$  activities to their equilibrium levels, effectively allowing their instantaneous equilibration and corresponding to  $\tau_n = 0$ .

We have also carried out extensive simulations with non-instantaneous  $n_i$ -dynamics and nonnegative  $n_i$ -values for more biological realism as per Eqs (8) and (9); the results were similar (not shown in this paper).

Other fast-dynamics parameters were as follows:  $\epsilon = 0.01$ ,  $\tau_f = 1$ ; integration by the forward Euler method with time step  $\delta t = 0.03$ . The matrix of level-1 familiarity cells defining the length of the activity vectors  $f_1$  was 91-by-136 so that the total number of familiarity neurons was one half that of pixels in the image, whereas the number of the level-1 novelty neurons was equal to the total number of pixels in the image, i.e., 24576. The radius of the circular receptive field (RF) within level 1, defining the non-zero row entries for (24'576-by-12'288 dimensional) synaptic matrix  $U_1$ , was equal to 4 (yielding  $\approx 50$  entries for each row, i.e., pixels in the RF of a familiarity neuron). For level 2, the procedure was exactly the same, except that the feedforward input to level 2 was provided by the  $f_1$ -values at the end of the fast dynamics, and these values were not filtered. The size of the second-level familiarity matrix was 64-by-96, and the RF radius defining the row entries of (12'288-by-6'144 dimensional) matrix  $U_2$  was again 4 (which implies that its effective RF radius in terms of the original image is  $\approx 9.0 = 3.5 * \sqrt{2} + 4$ : there will be on average 3.5 distances ( $\sqrt{2}$  in terms of original interpixel distance) between level-1 familiarity cells within the level-2 RF and the level-1 familiarity cells outermost in the level-2 RF will contribute their full radius in terms of the original image, i.e., 4).

### Synaptic plasticity

The entire learning session consisted of presenting the 1000 filtered images, randomly sampled in a total of  $T = 5000$  presentations. Following the equilibration of the fast neuronal dynamics during an image presentation, the synaptic weights were updated according to Eqs (6) and (7). The non-zero entries of the weight matrices  $U$  and  $V$  (corresponding to the RFs) were



initialized independently with a mean of 0 and standard deviation of  $\approx 0.0006$  for both levels 1 and 2. The learning rate for  $U$  was initialized to be  $\eta_0^U = 0.0001$  for level 1 and  $\eta_0^U = 0.0002$  for level 2. During the course of the image presentations these learning rates were gradually reduced to one-fifth of the original rates according to the schedule  $\eta_k = \eta_0/(1 + 4k/T)$  that defined the learning rates at the  $k^{\text{th}}$  presentation ( $k = 1 \dots T$ ).

## Performance measures

To quantify how good a reconstruction of an input image familiarity neurons  $f_i$  at each level would allow, we evaluated the distance  $\delta$  between the filtered image  $\tilde{I}$  (LGN output) and its reconstruction based on the activity of level-1 or level-2 familiarity neurons,  $\tilde{I} \approx U_1 f_1$  and  $\tilde{I} \approx U_1 U_2 f_2$ , respectively. As a distance measure between two images (or between two activity vectors)  $a$  and  $b$  we used the Euclidean ( $l^2$ ) norm of the difference of normalized images (represented as vectors),  $\delta(a, b) = \|a/\|a\| - b/\|b\|\|$ . For visual comparison purposes, we also presented the reconstructed images throughout the paper as inverted back into the original image space (e.g. [Fig 1d1 and 1d2](#)). The brain does not need to perform such an explicit reconstruction. Instead, it may differentially use the local familiarity and novelty information at each level for further processing and for obtaining the full original information. In fact, after the relaxation of the neuronal dynamics ([Eq \(4\)](#) with  $\dot{f}_i = \dot{n}_i = 0$ ), the lower-level activity could be fully reconstructed, e.g.  $f_0 = U_1 f_1 + n_1$ , while the information is segregated into familiarity and novelty representations at the higher level ([Fig 1c–1e](#)).

To trace the various quantities during the learning process in [Fig 3a](#), we low-pass filtered these quantities according to  $\bar{x} := (1 - \lambda)\bar{x} + \lambda x$ , with  $\lambda = 0.02$ ,  $x$  representing the reconstruction error  $\delta$  or the spatial average of neuronal activities  $f_1$  and  $n_1$ , respectively, and  $\bar{x}$  being the low-pass-filtered version of the same quantity.

The reconstruction error in [Fig 5c](#) was calculated using the distance measure  $\delta(a, b)$  between the reconstructed images  $a$  and  $b$  corresponding to cases with and without the occlusion, respectively. This choice highlights the effects of occlusion rather than reconstruction errors present both with and without the occlusion.

## Supporting Information

**S1 Supporting Information. Modular predictive coding analysis in further detail.**  
(PDF)

## Acknowledgments

This work was supported by the Swiss National Science Foundation (SNSF, personal grants No. 31003A-133094 and 310030L-156863 of WS).

## Author Contributions

Conceived and designed the experiments: BV WS RU. Performed the experiments: BV. Analyzed the data: BV WS RU. Wrote the paper: WS BV.

## References

1. Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. *Nat Neurosci.* 1999; 2(11):1019–1025. doi: [10.1038/14819](https://doi.org/10.1038/14819) PMID: [10526343](https://pubmed.ncbi.nlm.nih.gov/10526343/)
2. Cadieu C, Kouh M, Pasupathy A, Connor CE, Riesenhuber M, Poggio T. A model of V4 shape selectivity and invariance. *J Neurophys.* 2007; 98(3):1733–1750. doi: [10.1152/jn.01265.2006](https://doi.org/10.1152/jn.01265.2006)

3. Lichtman J, Livet J, Sanes J. A technicolour approach to the connectome. *Nat Neurosci.* 2008; 9 (6):417–422. doi: [10.1038/nm2391](https://doi.org/10.1038/nm2391)
4. Seung H. Reading the book of memory: sparse sampling versus dense mapping of connectomes. *Neuron.* 2009; 62(1):17–29. doi: [10.1016/j.neuron.2009.03.020](https://doi.org/10.1016/j.neuron.2009.03.020) PMID: [19376064](https://pubmed.ncbi.nlm.nih.gov/19376064/)
5. Friston K. Functional and effective connectivity in neuroimaging: a synthesis. *Human Brain Mapping.* 1994; 2:56–78. doi: [10.1002/hbm.460020107](https://doi.org/10.1002/hbm.460020107)
6. Koch M, Norris D, Hund-Georgiadis M. An investigation of functional and anatomical connectivity using magnetic resonance imaging. *Neuroimage.* 2002; 16(1):241–251. doi: [10.1006/nimg.2001.1052](https://doi.org/10.1006/nimg.2001.1052) PMID: [11969331](https://pubmed.ncbi.nlm.nih.gov/11969331/)
7. Rao R, Ballard D. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neurosci.* 1999; 2:79–87. doi: [10.1038/4580](https://doi.org/10.1038/4580)
8. Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. Canonical microcircuits for predictive coding. *Neuron.* 2012; 76(4):695–711.
9. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci.* 2013; 36:181–253. doi: [10.1017/S0140525X12000477](https://doi.org/10.1017/S0140525X12000477)
10. Koch C, Poggio T. Predicting the visual world: silence is golden. *Nature Neurosci.* 1999; 2:9–10. doi: [10.1038/4511](https://doi.org/10.1038/4511) PMID: [10195172](https://pubmed.ncbi.nlm.nih.gov/10195172/)
11. Spratling MW. Reconciling predictive coding and biased competition models of cortical function. *Front Comp Neurosci.* 2008; 2:4.
12. Hinton GE, Dayan P, Frey BJ, Neal RM. The “Wake-Sleep” Algorithm for Unsupervised Neural Networks. *Science.* 1995; 268:1158–1161. doi: [10.1126/science.7761831](https://doi.org/10.1126/science.7761831) PMID: [7761831](https://pubmed.ncbi.nlm.nih.gov/7761831/)
13. Dayan P, Hinton GE, Neal RM, Zemel RS. The Helmholtz machine. *Neur Comp.* 1995; 7:889–904. doi: [10.1162/neco.1995.7.5.889](https://doi.org/10.1162/neco.1995.7.5.889)
14. Barlow HB. Possible principles underlying the transformation of sensory messages. In: Rosenblith WA, editor. *Sensory Communication.* MIT Press; 1961. p. 217–234.
15. Barlow HB. Redundancy reduction revisited. *Network.* 2001; 12:241–253. doi: [10.1080/net.12.3.241.253](https://doi.org/10.1080/net.12.3.241.253) PMID: [11563528](https://pubmed.ncbi.nlm.nih.gov/11563528/)
16. Friston K. The free-energy principle: a unified brain theory? *Nature Rev Neurosci.* 2010; 11:127–138. doi: [10.1038/nrn2787](https://doi.org/10.1038/nrn2787)
17. Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. *Nature.* 1996; 381:520–522. doi: [10.1038/381520a0](https://doi.org/10.1038/381520a0) PMID: [8632824](https://pubmed.ncbi.nlm.nih.gov/8632824/)
18. Thorpe SJ. The speed of categorization in the human visual system. *Neuron.* 2009; 62(2):168–170. doi: [10.1016/j.neuron.2009.04.012](https://doi.org/10.1016/j.neuron.2009.04.012)
19. Stanford TR, Shankar S, Massoglia DP, Costello MG, Salinas E. Perceptual decision making in less than 30 milliseconds. *Nat Neurosci.* 2010; 13(3):379–385. doi: [10.1038/nn.2485](https://doi.org/10.1038/nn.2485) PMID: [20098418](https://pubmed.ncbi.nlm.nih.gov/20098418/)
20. Spratling MW. Predictive coding as a model of response properties in cortical area V1. *J Neurosci.* 2010; 30(9):3531–3543. doi: [10.1523/JNEUROSCI.4911-09.2010](https://doi.org/10.1523/JNEUROSCI.4911-09.2010) PMID: [20203213](https://pubmed.ncbi.nlm.nih.gov/20203213/)
21. Friston KJ. Hierarchical models in the brain. *PLoS Comp Biol.* 2008; 4:e1000211. doi: [10.1371/journal.pcbi.1000211](https://doi.org/10.1371/journal.pcbi.1000211)
22. Fyfe C. A neural network for PCA and beyond. *Neural Processing Letters.* 1997; 6:33–41. doi: [10.1023/A:1009606706736](https://doi.org/10.1023/A:1009606706736)
23. Oja E. The nonlinear PCA learning rule in independent component analysis. *Neurocomputing.* 1997; 17 (1):25–45. doi: [10.1016/S0925-2312\(97\)00045-3](https://doi.org/10.1016/S0925-2312(97)00045-3)
24. Plumbley MD, Oja E. A “nonnegative PCA” algorithm for independent component analysis. *IEEE Trans Neural Netw.* 2004; 15(1):66–76. doi: [10.1109/TNN.2003.820672](https://doi.org/10.1109/TNN.2003.820672) PMID: [15387248](https://pubmed.ncbi.nlm.nih.gov/15387248/)
25. Foldiak P. Adaptive network for optimal linear feature extraction. In: *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks.* vol. 1. IEEE Press; 1989. p. 401–405.
26. Haykin SO. *Neural Networks and Learning Machines.* 3rd ed. Prentice Hall; 2008.
27. Olshausen BA, Field DJ. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res.* 1997; 37(23):3311–3325. doi: [10.1016/S0042-6989\(97\)00169-7](https://doi.org/10.1016/S0042-6989(97)00169-7) PMID: [9425546](https://pubmed.ncbi.nlm.nih.gov/9425546/)
28. Hoyer PO. Modeling receptive fields with non-negative sparse coding. *Neurocomputing.* 2003; 52–54:247–252.
29. Lee H, Battle A, Raina R, Ng AY. Efficient sparse coding algorithms. *Adv Neural Inf Process Syst (NIPS).* 2006; 19:801–808.
30. Yeh CI, Xing D, Williams PE, Shapley RM. Stimulus ensemble and cortical layer determine V1 spatial receptive fields. *PNAS.* 2009; 106(34):14652–14657. doi: [10.1073/pnas.0907406106](https://doi.org/10.1073/pnas.0907406106) PMID: [19706551](https://pubmed.ncbi.nlm.nih.gov/19706551/)

31. Victor JD, Mechler F, Ohiorhenuan I, Schmid AM, Purpura KP. Laminal and orientation-dependent characteristics of spatial nonlinearities: implications for the computational architecture of visual cortex. *J Neurophysiol*. 2009; 102:3414–3432. doi: [10.1152/jn.00086.2009](https://doi.org/10.1152/jn.00086.2009) PMID: [19812295](https://pubmed.ncbi.nlm.nih.gov/19812295/)
32. Fournier J, Monier C, Pananceau M, Fregnac Y. Adaptation of the simple or complex nature of V1 receptive fields to visual statistics. *Nat Neurosci*. 2011; 14(8):1053–1060. doi: [10.1038/nn.2861](https://doi.org/10.1038/nn.2861) PMID: [21765424](https://pubmed.ncbi.nlm.nih.gov/21765424/)
33. Bolz J, Gilbert C. Generation of end-inhibition in the visual cortex via interlaminar connections. *Nature*. 1986; 320:362–365. doi: [10.1038/320362a0](https://doi.org/10.1038/320362a0) PMID: [3960119](https://pubmed.ncbi.nlm.nih.gov/3960119/)
34. Durand S, Freeman TC, Carandini M. Temporal properties of surround suppression in cat primary visual cortex. *Vis Neurosci*. 2007; 24(5):679–690. doi: [10.1017/S0952523807070563](https://doi.org/10.1017/S0952523807070563) PMID: [17686200](https://pubmed.ncbi.nlm.nih.gov/17686200/)
35. Chavane F, Sharon D, Jancke D, Marre O, Fregnac Y, Grinvald A. Lateral Spread of Orientation Selectivity in V1 is Controlled by Intracortical Cooperativity. *Front Syst Neurosci*. 2011; 5:4. doi: [10.3389/fnsys.2011.00004](https://doi.org/10.3389/fnsys.2011.00004) PMID: [21629708](https://pubmed.ncbi.nlm.nih.gov/21629708/)
36. Friston KJ, Kiebel S. Predictive coding under the free-energy principle. *Philos Trans R Soc Lond, B, Biol Sci*. 2009; 364:1211–1221. doi: [10.1098/rstb.2008.0300](https://doi.org/10.1098/rstb.2008.0300) PMID: [19528002](https://pubmed.ncbi.nlm.nih.gov/19528002/)
37. Ohzawa I, DeAngelis G, Freeman R. Encoding of binocular disparity by complex cells in the cat's visual cortex. *J Neurophys*. 1997; 77:2879–2909.
38. Adelson EH, Bergen JR. Spatiotemporal energy models for the perception of motion. *J Opt Soc Am A*. 1985; 2(2):284–299. doi: [10.1364/JOSAA.2.000284](https://doi.org/10.1364/JOSAA.2.000284) PMID: [3973762](https://pubmed.ncbi.nlm.nih.gov/3973762/)
39. Miller K. A model for the development of simple cell receptive field and the ordered arrangement of orientation columns through activity-dependent competition between ON- and OFF-center input. *J Neurophys*. 1994; 14:409–441.
40. Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*. 1996; 381:607–609. doi: [10.1038/381607a0](https://doi.org/10.1038/381607a0) PMID: [8637596](https://pubmed.ncbi.nlm.nih.gov/8637596/)
41. Karklin Y, Lewicki MS. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*. 2009; 457:83–86. doi: [10.1038/nature07481](https://doi.org/10.1038/nature07481) PMID: [19020501](https://pubmed.ncbi.nlm.nih.gov/19020501/)
42. Larkum M, Senn W, Lüscher HR. Top-down dendritic input increases the gain of layer 5 pyramidal neurons. *Cereb Cortex*. 2004; 14:1059–1070. doi: [10.1093/cercor/bhh065](https://doi.org/10.1093/cercor/bhh065) PMID: [15115747](https://pubmed.ncbi.nlm.nih.gov/15115747/)
43. Johnson R, Burkhalter A. A polysynaptic feedback circuit in rat visual cortex. *J Neurosci*. 1997; 17(18):7129–7140.
44. Angelucci A, Bullier J. Reaching beyond the classical receptive field of V1 neurons: horizontal or feedback axons? *J Physiol (Paris)*. 2003; 97:141–154. doi: [10.1016/j.jphysparis.2003.09.001](https://doi.org/10.1016/j.jphysparis.2003.09.001)
45. Cohen MR, Maunsell JH. Using neuronal populations to study the mechanisms underlying spatial and feature attention. *Neuron*. 2011; 70(6):1192–1204. doi: [10.1016/j.neuron.2011.04.029](https://doi.org/10.1016/j.neuron.2011.04.029) PMID: [21689604](https://pubmed.ncbi.nlm.nih.gov/21689604/)
46. Nyberg L, Habib R, McIntosh AR, Tulving E. Reactivation of encoding-related brain activity during memory retrieval. *Proc Natl Acad Sci USA*. 2000; 97(20):11120–11124. doi: [10.1073/pnas.97.20.11120](https://doi.org/10.1073/pnas.97.20.11120)
47. Garrido MI, Kilner JM, Kiebel SJ, Stephan KE, Baldeweg T, Friston KJ. Repetition suppression and plasticity in the human brain. *Neuroimage*. 2009; 48(1):269–279. doi: [10.1016/j.neuroimage.2009.06.034](https://doi.org/10.1016/j.neuroimage.2009.06.034) PMID: [19540921](https://pubmed.ncbi.nlm.nih.gov/19540921/)
48. Lieder F, Stephan KE, Daunizeau J, Garrido MI, Friston KJ. A neurocomputational model of the mismatch negativity. *PLoS Comput Biol*. 2013; 9(11):e1003288. doi: [10.1371/journal.pcbi.1003288](https://doi.org/10.1371/journal.pcbi.1003288) PMID: [24244118](https://pubmed.ncbi.nlm.nih.gov/24244118/)
49. Urbanczik R, Senn W. Learning by the dendritic prediction of somatic spiking. *Neuron*. 2014; 81(3):521–528. doi: [10.1016/j.neuron.2013.11.030](https://doi.org/10.1016/j.neuron.2013.11.030) PMID: [24507189](https://pubmed.ncbi.nlm.nih.gov/24507189/)
50. Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, et al. Do we know what the early visual system does? *J Neurosci*. 2005; 25(46):10577–10597. doi: [10.1523/JNEUROSCI.3726-05.2005](https://doi.org/10.1523/JNEUROSCI.3726-05.2005) PMID: [16291931](https://pubmed.ncbi.nlm.nih.gov/16291931/)
51. Van Hateren J, Van Der Schaaf A. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc R Soc Lond, B, Biol Sci*. 1998; 265:359–366.

---

# Supporting Information

for

## Hierarchical Novelty-familiarity Representation in the Visual System by Modular Predictive Coding

Boris Vladimirovskiy<sup>1,□a</sup>, Robert Urbanczik<sup>1</sup>, Walter Senn<sup>1,\*</sup>

**1 Department of Physiology, University of Bern, Bülhplatz 5, 3012 Bern, Switzerland**

**□a Currently at Munich Re UK & Ireland Life, Plantation Place, 30 Fenchurch Street, London EC3M 3AJ, United Kingdom**

**\* Senn@cns.physio.unibe.ch**

Here we first present the framework of the classical ‘cross-level’ predictive coding (Section S.I). We show that for a single level, with a readout at the next level but without top-down connectivity, predictive coding performs principal component analysis (Section S.II). We next argue that cross-level predictive coding across 2 levels can essentially be reduced to a single level (Section S.III). To keep the computational benefit of a hierarchical principal component analysis and to reduce processing time throughout the levels, we therefore reduce the cross-level predictive coding to our modular predictive coding scheme (Section S.IV). These sections form the background for Subsection of the main text in which we consider alternative coding schemes and conclude that in terms of a neuronal implementation, all the generative models (e.g., as reviewed in [1]) essentially reduce to the quadratic error function with a linear transfer function.

### S.I Classical cross-level predictive coding

In [2], predictive coding was introduced as a hierarchical model of visual processing based on a global objective function summed across the individual components

$$E_i = E_{U_i}(f_i, f_{i-1}) = \frac{1}{2} \|f_{i-1} - \phi(U_i f_i)\|^2 \quad (\text{S.1})$$

at each level  $i$  of the hierarchy (cf. Eq. (1) in the main text). Here,  $f_i$  is the vector of neuronal activations at level  $i$  (with ‘ $f$ ’ standing for ‘familiarity’), and  $f_0 = I$  being the input into the system.  $U_i$  is a matrix of suitable dimensions with  $\phi(U_i f_i)$  representing the prediction of the activity in level  $i - 1$  based on level- $i$  activity.

Notice that in a Gaussian noise scenario, where each component of the low level activity  $f_{i-1}$  is contaminated by an additive Gaussian noise of mean 0 and variance 1, the log-likelihood of the ‘data’  $f_{i-1}$  given the ‘parameters’  $U_i$  and  $f_i$  is just a constant minus the error function (Eq. (S.1)), i.e., formally  $\log P(f_{i-1} | U_i, f_i) = C - E_i$ , with  $C = -N_{i-1} \log(2\pi)/2$  and  $N_{i-1}$  the dimension of  $f_{i-1}$ . Hence, maximizing the log-likelihood of the lower-level data is equivalent to minimizing the reconstruction error. This links the predictive coding scheme (S.1) to various types of Bayesian modeling [1].

The global objective function according to [2] is just a weighted sum over the levels,

$$E_1 + \dots + E_L = E(U, f; I) = \sum_{i=1}^L E_{U_i}(f_i, f_{i-1}). \quad (\text{S.2})$$

Recognition in this system amounts to minimizing  $E(U, f; I)$  with respect to  $f_i$  for  $i > 0$  to obtain

$$E(U; I) = \min_{f_i, i>0} E(U, f; I),$$

and learning amounts to a slow dynamics in  $U_i$  minimizing the average  $\langle E(U; I) \rangle_I$  of the errors obtained during recognition.

For recognition the relevant gradients defining the temporal evolution  $df_i/dt$  of the neuronal activity in the  $i$ -th level are

$$\nabla_{f_i} E(U, f; I) = \nabla_{f_i} E_{U_i}(f_i, f_{i-1}) + \nabla_{f_i} E_{U_{i+1}}(f_{i+1}, f_i) \quad (\text{S.3})$$

if  $1 \leq i < L$ , whereas for  $i = L$  the second term in the sum is absent. Hence the gradient in  $f_i$  also depends on the activations in level  $i + 1$  and this in effect introduces a global coupling of the system. If there are many levels such a global relaxation process becomes difficult to reconcile with the time constraints of visual processing.

## S.II Single-level predictive coding performs principle component analysis (PCA)

Since, as per Subsection 2.7 of the main text, only linear generative functions  $\phi$  have a neuronal implementation we set  $\phi = \text{id}$  (as also done in [2]). We first consider a single level  $L = 1$  so that the global objective function reduces to

$$E_1 = E_{U_1}(f_1, I) = \frac{1}{2} \|I - U_1 f_1\|^2$$

Minimizing this with respect to  $f_1$  yields  $f_1 = U_1^+ I$ , where  $U_1^+$  is the Moore-Penrose pseudo-inverse<sup>1</sup> of  $U_1$ . Inserting this in the above inequality yields for  $E(U; I) = \frac{1}{2} \|(1 - U_1 U_1^+) I\|^2$ . Note that  $U_1 U_1^+ = P_{U_1}$  represents the orthonormal projection onto the subspace spanned by the columns of  $U_1$ . Similarly,  $1 - P_{U_1} = P_{U_1^\perp}$  represents the orthonormal projection onto the orthogonal complement of the space spanned by the columns of  $U_1$ . We hence obtain  $E(U_1; I) = \frac{1}{2} \|P_{U_1^\perp} I\|^2$ .

Using the last equation we calculate the expected error across images:

$$\begin{aligned} \langle E(U_1; I) \rangle_I &= \frac{1}{2} \left\langle \left( P_{U_1^\perp} I \right)^T P_{U_1^\perp} I \right\rangle_I \\ &= \frac{1}{2} \left\langle \text{Tr} \left( P_{U_1^\perp} I \left( P_{U_1^\perp} I \right)^T \right) \right\rangle_I \\ &= \frac{1}{2} \left\langle \text{Tr} \left( P_{U_1^\perp} I I^T \left( P_{U_1^\perp} \right)^T \right) \right\rangle_I \\ &= \frac{1}{2} \text{Tr} \left( P_{U_1^\perp} \langle I I^T \rangle_I \right). \end{aligned}$$

To arrive at the last line we used the fact that within the argument of the trace operator the three symmetric matrices commute,  $\text{Tr}(ABC) = \text{Tr}(ACB)$ , and that for any projection  $P_U$  one has  $P_U P_U^T = P_U^2 = P_U$ . Hence, to find the minimum of the above expression across  $U_1$ , the columns of  $U_1^\perp$  must span the subspace of eigenvectors

<sup>1</sup>If the columns of  $U = U_1$  are linearly independent then  $U^T U$  is invertible and  $U^+ = (U^T U)^{-1} U^T$ .

of the correlation matrix  $\langle II^T \rangle_I$  with the smallest eigenvalue. Put differently, the columns of the matrix  $U_1$  itself must span the eigenspace corresponding to the largest eigenvalues. If  $U_1$  has  $N_1$  columns and an image  $I$  is encoded by  $N_0$  pixels, the dimension of this subspace is  $N_0 - N_1$ . We then get

$$\min_{U_1} \langle E(U_1; I) \rangle_I = \frac{1}{2} \min_{U_1} \text{Tr} \left( P_{U_1^\perp} \langle II^T \rangle_I \right) = \frac{1}{2} \sum_{k=N_1+1}^{N_0} \lambda_k^{(0)}, \quad (\text{S.4})$$

where  $\lambda_k^{(0)}$  are the  $N_0 - N_1$  smallest eigenvalues of the correlation matrix  $\langle II^T \rangle_I$ .

Hence,  $U_1 U_1^+ = P_{U_1}$  is the projection onto the first  $N_1$  principal components of the images  $I$ , and since in the steady state we have  $f_1 = U_1^+ I$ , the  $f_1$  must jointly encode these first principal components. Thus, finding  $U_1$  which minimizes the energy function essentially corresponds to doing principal component analysis (PCA). Compared to other PCA algorithms, the present procedure may be more neural but is also less constrained, since, after minimizing in  $f_1$ , the relevant object is not  $U_1$  but  $P_{U_1}$ . So any two matrices spanning the same subspace become equivalent and, in particular, there is no reason why the  $U_1$  we find should have, say, orthonormal columns. For the same reason there is no scale in the system, any nonzero scalar multiple of  $U_1$  is just as good as  $U_1$ .

In many applications of predictive coding, the possible values of  $U_1$  are constrained by the fact that the neurons of  $f_1$  should have limited receptive fields (RFs). In fact, if a RF of a first level neuron covers  $p_0$  pixels in the image, the corresponding column of  $U_1$  will have  $N_0 - p_0$  zero entries. However, this does not affect the above analysis, and there the full optimization just has to be replaced by a constrained optimization. This leads to a constrained PCA where the individual neurons jointly represent the eigenspaces of the input. Assuming an  $n_1$ -fold overlap of RFs at any position in the image space, the minimization procedure will extract the first  $n_1$  eigenvectors of the  $N_0 \times N_0$  correlation matrix  $\langle II^T \rangle_I$  averaged across all images  $I$ .

### S.III Reduction of cross-level to modular predictive coding

#### Error minimization across neighboring levels

$$E(U, f; I) = E_{U_1}(f_1, I) + E_{U_2}(f_2, f_1).$$

Minimizing  $E(U, f; I)$  across  $f_1$  and  $f_2$  will, at the top level, still lead to an optimal  $f_2$  of the form  $f_2 = U_2^+ f_1$ , whatever  $f_1$  is. As above we calculate

$$\begin{aligned} E_{U_2}(f_2, f_1) &= \frac{1}{2} \|f_1 - U_2 f_2\|^2 = \frac{1}{2} \|(1 - P_{U_2})f_1\|^2 \\ &= \frac{1}{2} \|P_{U_2^\perp} f_1\|^2. \end{aligned}$$

Note that this reduces the optimization problem across both levels to an optimization within only the first level,

$$E(U, f; I) = \frac{1}{2} \|I - U_1 f_1\|^2 + \frac{1}{2} \|P_{U_2^\perp} f_1\|^2. \quad (\text{S.5})$$

The reduced optimization problem consists of finding the  $N_1 \times N_2$  matrix  $U_1$  (mapping the  $N_2$ - to the  $N_1$ -dimensional space) and the  $(N_1 - N_2)$ -dimensional subspace orthogonal to the columns of  $U_2$  within the  $N_1$ -dimensional space.

We next plug in as  $f_1$  the activity  $f_1^*$  which minimizes the total error and take the average across the images  $I$ . Applying again the calculations underlying Eq. S.4 yields

the estimate

$$\begin{aligned}\langle E(U; I) \rangle_I &= \langle E_{U_1}(f_1^*, I) \rangle_I + E_2, \\ E_2 &\geq \frac{1}{2} \sum_{k=N_2+1}^{N_1} \lambda_k^{(1)},\end{aligned}\tag{S.6}$$

where  $\lambda_i^{(1)}$ 's represent the  $N_1 - N_2$  smallest eigenvalues of the correlation matrix  $\langle f_1^* f_1^{*T} \rangle_I$ . For the first level we have that  $E_{U_1}(f_1^*, I) \geq E_{U_1}(U_1^+ I, I)$ , so the fact that  $f_1^*$  can depend on  $U_2$  does not help to make the first term in (S.6) smaller and hence

$$\langle E_{U_1}(f_1^*, I) \rangle_I \geq \frac{1}{2} \sum_{k=N_1+1}^{N_0} \lambda_k^{(0)}.$$

In contrast to the single-level optimization, however, it is not clear that the first level lower bound is attained when optimizing the objective function summed across both levels, since this might yield a choice of  $f_1$  which extracts a high penalty in  $E_2$ . So in the joint optimization there is a trade-off between the prediction quality in the first level on one hand, and obtaining an easily predictable distribution of  $f_1$  for the second level on the other hand. But if nature does not use the full expressive power provided by the  $N_1$  neurons in level 1, it might quite simply choose a smaller value of  $N_1$  and, instead, use additional neurons to interpret the activity in the second level. From this efficiency perspective, the advantage of miming the sum  $E_1 + E_2$  over minimizing  $E_1$  and  $E_2$  individually remains therefore unclear. In contrast, the individual minimization has a distinct advantage that is particularly useful in the sensory representation and that goes beyond the processing speed argument: when information at the different resolution levels needs to be accessed, it is individually optimized at the level of interest. The joint minimization instead may obscure the representation at that level, for the sake optimizing it at the next level.

## Reduction to a single level

Given the unclear motivation to consider top-down interactions in terms of image encoding we ask whether we cannot compensate for the top-down signaling by enlarging the interactions within the lower level. This is in fact possible, as alluded to in (S.5). To make this point clear, we express  $f_1$  in new coordinates  $\tilde{f}$  (specified below) such that the second term in (S.5) reduces to a sum over squared activities. The 2-layer optimization problem (Eq. S.2 with  $L = 2$ ) then reduces via (S.5) to

$$E(\tilde{U}, \tilde{f}; I) = \frac{1}{2} \|I - \tilde{U}_1 \tilde{f}\|^2 + \frac{1}{2} \sum_{k=N_2+1}^{N_1} (\tilde{f}_k)^2,\tag{S.7}$$

where  $N_1$  and  $N_2$  are the number of neurons in the first and second level, respectively.

Recognition again amounts to the minimization of (S.7) with respect to the activities  $\tilde{f}_1$  of the first-level neurons, while learning amounts to the minimization of (S.7) with respect to  $\tilde{U}_1$  at the minimum values of  $\tilde{f}_1$ . The additional constraint imposed by the joint 2-level minimization is on the mapping  $\tilde{U}_1$  that emerges from the specific coordinate transform (see next paragraph). Basically, (S.7) states that the 2-level error function  $E_1 + E_2$  is minimized by finding a representation  $f_1$  at the first level that minimizes  $E_1$  while keeping small the  $(N_1 - N_2)$  components that are not represented at the second level. Of course, when constructing the first level representation it is not clear which dimensions will be cut off in the second level representations. But instead of

evoking top-down connections from higher levels that would deliver this information – and would also slow down the network relaxation time – we explored an effectively feedforward strategy.

One strategy to mimic the top-down effect in (S.7) is to uniformly keep the  $f_1$  activities small for all components and defer the compression of the  $f_1$ -code to the next level. We therefore suggest to directly minimize the quadratic error function  $E_i$  at each level  $i$ , with an additional regularity constraint of the form  $\sum_k (\tilde{f}_i)_k^2$ , where the sum now extends across all  $N_i$  components of  $\tilde{f}_i$ , i.e.

$$E_i = \frac{1}{2} \|I - U_i f_i\|^2 + \frac{\epsilon}{2} \sum_{k=1}^{N_i} (f_i)_k^2, \quad (\text{S.8})$$

although with an  $\epsilon$  smaller than 1/2. This regularity limits the size of the  $f_i$  activity, but it does not impose sparseness, as e.g. the L1-norm on  $f_i$  would do [3]. We did not consider specific sparseness constraints by several reasons. First, the experimental notion of sparseness is somewhat vague and in biology firing rates are often small but rarely zero, as it would be the case for sparse coding. The square norm in (S.8) impedes high firing rates of individual neurons but allows for uniformly low firing rates across the network. Second, sparseness would lead to strict Gabor-like RFs [3] that, for non-grating stimuli, do not seem to square with the context-dependent RF variations [4–7]. Third, from an economical point of view, it is more efficient to achieve a high stimulus compression rate by limiting the number of  $f_i$  neurons, rather by enforcing them mostly not to respond. Finally, abstaining from additional penalty terms increases the clarity of the procedure.

The coordinate transform leading to the form (S.7) is given by  $f_1 = S \tilde{f}_1$ , where the columns of the orthogonal matrix  $S$  are the normalized eigenvectors of the correlation matrix  $\langle f_1^* f_1^{*T} \rangle_I$ , ordered according to decreasing eigenvalues (assumed to be non-vanishing). Notice that  $S$  can be written as a concatenation  $S = U_2 \oplus U_2^\perp$  of a  $N_1 \times N_2$  dimensional matrix  $U_2$  and a  $N_1 \times (N_1 - N_2)$  dimensional matrix  $U_2^\perp$ . Note that the projection matrix  $P_{U_2^\perp}$  arising in (S.5) transforms into the diagonal matrix  $S^T P_{U_2^\perp} S$  with 0's in the first  $N_2$  entries and 1's in the  $N_1 - N_2$  remaining entries. With  $\tilde{U}_1 = U_1 S$  this proves (S.7). The effective receptive fields of the  $f_1$  neurons are given by the non-vanishing row entries of  $\tilde{U}_1$ , and via the correlation matrix  $S$  they depend on the overlaps of the feedforward receptive fields of  $f_1$  and, in the case of the fully-recurrent network, they also depend on the top-down input from the higher level.

## S.IV Implementation of modular and cross-level hierarchical coding

The above analysis suggests decoupling the optimization problem for multiple levels, thus enhancing the clarity of the procedure as well as increasing the speed of convergence. Hence, instead of minimizing the overall error across levels (Eq. S.2), we individually minimize the errors (S.1) within each level. For the (negative) gradients with respect to  $f_i$  and  $U_i$  with  $\phi = \text{id}$  we obtain for the modular case,

$$\begin{aligned} \tau \dot{f}_i &= -\epsilon f_i + V_i n_i \\ \tau \dot{n}_i &= -n_i + f_{i-1} - U_i f_i \\ \Delta U_i &= \eta n_i^* (f_i^*)^T \end{aligned} \quad (\text{S.9})$$

with  $V_i = U_i^T$  and  $\Delta V_i = \Delta U_i^T$  (cf. Eqs. (4) and (6) in the main text).



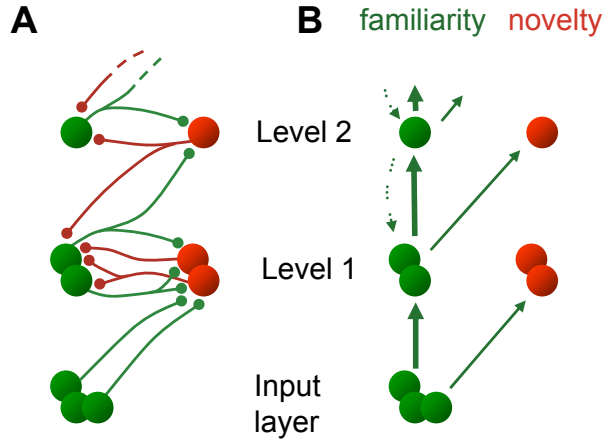
When considering the original, non-modular optimization principle where minimization is performed across all levels as in [2], we need to implement the gradient (S.3). This leads to the same three equations as in (S.9), except the first equation is replaced by

$$\tau \dot{f}_i = -\epsilon f_i + V_i n_i - n_{i+1}, \quad (\text{S.10})$$

and hence during the learning process  $f_i$ 's are implicitly modified until the novelty  $n_{i+1}$  at the higher level cannot be explained anymore by combining the novelty  $n_i$  at level  $i$  via matrix  $V_i$  (and assuming that  $\epsilon$  is small, cf. also Fig. Aa). At a functional level this corresponds to the constrained optimization problem for  $f_i$  as expressed in Eqs. S.5 and S.7, without the need to consider the activities from the next level  $i + 1$ . An equivalent implementation of the cross-level optimization problem following Eq. S.7 leads to the same dynamics as in (S.9), but with the first equation replaced by a neuronal dynamics with weak and strong neuronal leak terms,

$$\tau (\dot{f}_i)_k = \begin{cases} -\epsilon (f_i)_k + (V_i n_i)_k & \text{for } k = 1, \dots, N_{i+1} \\ -(1 + \epsilon)(f_i)_k + (V_i n_i)_k & \text{for } k = N_{i+1} + 1, \dots, N_i, \end{cases} \quad (\text{S.11})$$

where  $N_i$  and  $N_{i+1}$  are the numbers of familiarity neurons at level  $i$  and  $i + 1$ , respectively. Thus, recognition is again performed in an effectively feedforward network, with simple (i.e. diagonalized Gaussian) activity priors depending on the number of neurons in the  $i + 1$ 'th level (Fig. Ab).



**Figure A.** Anatomical versus effective connectivity for the cross-level predictive coding as in [2]. **A** In contrast to minimizing the error function locally at each level, minimizing the summed error function across all levels requires top-down connections from the novelty neurons  $n_{i+1}$  to the familiarity neurons  $f_i$  (cf. Eqs. (S.9), (S.10) and Fig. (2) in the main text). **B** The effective connectivity expressing causal relationships remains feedforward. The only ‘top-down’ information (dotted arrows) consists in providing the number of familiarity neurons used at level  $i + 1$ , and this determines the number of familiarity neurons at level  $i$  which are subject to a strong leak in the neuronal dynamics (cf. Eqs. (S.7) and (S.11)). This distortion caused by the strong leak may reduce the prediction error  $E_{i+1}$  at the upper level, while possibly increasing the error  $E_i$  at the lower level (with an overall reduction in the sum  $E_i + E_{i+1}$ ).

---

## References

1. Friston K. The free-energy principle: a unified brain theory? *Nature Rev Neurosci.* 2010;11:127–138.
2. Rao R, Ballard D. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neurosci.* 1999;2:79–87.
3. Olshausen BA, Field DJ. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res.* 1997;37(23):3311–3325.
4. Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, et al. Do we know what the early visual system does? *J Neurosci.* 2005;25(46):10577–10597.
5. Yeh CI, Xing D, Williams PE, Shapley RM. Stimulus ensemble and cortical layer determine V1 spatial receptive fields. *PNAS.* 2009;106(34):14652–14657.
6. Victor JD, Mechler F, Ohiorhenuan I, Schmid AM, Purpura KP. Laminar and orientation-dependent characteristics of spatial nonlinearities: implications for the computational architecture of visual cortex. *J Neurophysiol.* 2009;102:3414–3432.
7. Fournier J, Monier C, Pananceau M, Fregnac Y. Adaptation of the simple or complex nature of V1 receptive fields to visual statistics. *Nat Neurosci.* 2011;14(8):1053–1060.